Data Collection and Preprocessing Phase

Date	8th July 2024
Team ID	SWTID1719999219
Project Title	Crystal Clear Vision: Revolutionizing Cataract Prediction through Transfer Learning Mastery
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan

Section	Description
Project Overview	Crystal Clear Vision is on a mission to revolutionize early detection of cataracts, a major cause of blindness globally. Their approach utilizes cutting-edge deep learning, specifically transfer learning, to achieve this. Transfer learning allows us to harness the power of pre-existing models trained on vast datasets, ultimately aiming to develop a highly accurate and efficient system for predicting cataracts.
Data Collection Plan	● Search for datasets related to retinal fundus images, ophthalmological diagnoses, and cataract grading. ● Prioritize datasets with diverse patient demographics (age, ethnicity, etc.).

	Data collection is from popular open-source sites such as kaggle.com, Github and UCl repository. These will be explored to identify suitable datasets for our analysis.
Raw Data Sources Identified	Three raw data sources were identified for this project- 1. Ocular Disease Recognition dataset Kaggle 2. Retina dataset Github 3. Cataract dataset Kaggle

Raw Data Sources

Source Name	Description	Location/URL	Format	Size	Access Permissions
1. Ocular Disease Recognition dataset Kaggle	The "Ocular Disease Recognition" dataset has been taken up from Kaggle. It has images of around 5000 patients with age, color fundus photographs from left and right eyes and the doctor's	https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k	jpg files	2 GB	Public

	diagnostic keywords.				
Retina dataset Github	Retina dataset contains four categories: 1) normal 2) cataract 3) glaucoma 4) retina disease.	https://github.com/yiweichen04/retina_dataset	png	3.34 GB	Public
Cataract dataset Kaggle	Cataract dataset from Kaggle has two folders- train and test, each containing normal eye and cataract eye images.	https://www.kaggle.com/datasets/hemooredaoo/cataract	png files	579 MB	Public

The dataset which was finally selected for the project is Retina dataset from Github owing to its variety of images.